W208–W212 *Nucleic Acids Research, 2004, Vol. 32, Web Server issue*
DOI: 10.1093/nar/gkh474

SITECON: a tool for detecting conservative conformational and physicochemical properties in transcription factor binding site alignments and for site recognition

D. Y. Oshchepkov*, E. E. Vityaev, D. A. Grigorovich, E. V. Ignatieva and T. M. Khlebodarova

Institute of Cytology and Genetics SB RAS, Novosibirsk 630090, Russia

Received February 14, 2004; Revised April 21, 2004; Accepted May 3, 2004

ABSTRACT

The local DNA conformation in the region of transcription factor binding sites, determined by context, is one of the factors underlying the specificity of DNA–protein interactions. Analysis of the local conformation of a set of functional DNA sequences may allow for determination of the conservative conformational and physicochemical parameters reflecting molecular mechanisms of interaction. The web resource SITECON is designed to detect conservative conformational and physicochemical properties in transcription factor binding sites, contains a knowledge base of conservative properties for >100 high-quality sample sites and allows for recognition of potential transcription factor binding sites based on conservative properties from both the knowledge base and the results of analysis of a sample proposed by a user. The resource SITECON is available at http://wwwmgs.bionet.nsc.ru/mgs/programs/sitecon/.

INTRODUCTION

Specific binding of transcription factors to a DNA sequence is one of the key issues in understanding the fundamentals of transcription regulation. Statistical analysis of sample transcription factor binding sites allows common contextual characteristics used for recognition of potential sites to be detected. However, the data on context-dependent conformational and physicochemical properties may also be effectively used for both the analysis and recognition of transcription factor binding sites, since the local conformation of sites plays a role in recognition of a binding site by a transcription factor (1). The dependence of DNA conformation on context was first discovered by Dickerson and Drew in 1981 (2) by X-ray structure analysis of DNA 12mers. A growing volume of data from structural analyses demonstrated the non-uniformity of conformational and physicochemical properties and their dependence on nucleotide context (3,4).

When analyzing sample transcription factor binding sites, it is possible to find a context-dependent conformational or physicochemical property for each particular position of the alignment whose values for the variants of sites in the alignment at this position are very close. Moreover, despite the distinctions in context, dispersion of these values is considerably lower than for a random sample of sequences. As a rule, such low variation of properties can be found for a number of positions in the alignment of transcription factor binding sites. This is a result of the fact that sequences displaying a corresponding set of fixed values of conformational and physicochemical properties at certain positions allow a specific protein to bind with higher affinity. In other words, the local conformation of DNA molecules, determined by context, is a factor in the specificity of DNA–protein recognition (5). Such properties with close values in all the variants of sites of a sample differing in context will be detected by a directed search for low variation of properties among sites, which will reveal their low dispersion in particular positions of the sample.

Thus, certain properties at particular regions of the site should have fixed values to provide a successful binding between a DNA site and a specific protein. First and foremost, this is determined by the specificity of mechanisms of DNA–protein interactions for a particular DNA–protein complex (6). Thus, a complete set of data on conservative conformational and physicochemical properties of sites reflects the specificity of DNA sequence interaction with a particular protein and may be efficaciously used for recognition of potential binding sites.

Analysis of DNA properties has been successfully used for recognition (7) and analysis (8) of transcription factor binding sites. For example, Liu *et al.* (7) demonstrated, using the example of the MetJ transcription factor, that taking into account the properties of the DNA helix along with using

*To whom correspondence should be addressed. Tel: +7 3832 333119; Fax: +7 3832 331271; Email: diman@bionet.nsc.ru

traditional methods based on contextual analysis might increase the recognition quality. Earlier, Ponomarenko *et al.* (8) demonstrated that the activity of site DNA depended on a set of conformational parameters, which allowed the construction of several methods for determination of the activity of site DNA and for site recognition.

The web resource SITECON, described here, allows the detection of conservative properties of transcription factor binding sites based on statistical analysis of site samples and data on 38 conformational and physicochemical DNA properties. SITECON depicts on a colored map both the degree of conservation and the deviation of the mean values of each property at the positions of the sample, providing an efficient comparison of the properties of the sample with random sequences. Based on these data, SITECON also allows for constructing recognition rules and searching for potential transcription factor binding sites in genomic sequences. This resource also contains a knowledge base of conservative conformational and physicochemical properties for >100 transcription factors. Along with the sample transcription factor binding sites proposed by SITECON, users may analyze their own data in the form of an aligned sample of transcription factor binding sites. The method proposed has been successfully applied to recognition of some other transcription factor binding sites: IRFs, ISGF3, STATs, NF-κB (9), COUP-TF, PPRE (10), HSF (11) and E2F (12).

Now, several dozen conformational and physicochemical properties of the DNA helix have been determined. Characteristic of all these properties is their strong mutual correlation. However, it is necessary to make use of the maximum volume of readily interpretable data on particular conservative properties when analyzing data on the properties influencing the specificity of binding and the formation of a DNA–protein complex. Consequently, we use an exhaustive (in our opinion) set of 38 conformational and physicochemical DNA properties from the database Property (http://wwwmgs.bionet.nsc.ru/mgs/gnw/bdna/) (13).

METHODS

The essence of our approach is as follows. A training sample of aligned functional sites is used for detection of significantly conservative properties at the positions of alignment. Successively recoding the sequences into one of the 38 properties, we calculate the mean square deviation of the property at a the positions and the mean value. It is assumed that if a particular property at a particular location within the nucleotide sequence is important for the function of the binding site, the value of this property is conserved for all the sequences of the sample, providing a low value of the variance compared with a set of random sequences. Thus, a low variance of a particular property indicates its conservation at a particular position. The significance of the mean square deviation is estimated using chi-square test (14). These data are used to construct a table and the map of conservative properties. Then, comparing the properties of a query sequence with the conservative properties of the training sample, the total conformational similarity score is calculated. This score, is a measure of similarity between the sequences of the training sample and the sequence analyzed; in other words, this value is considered to be a 'score' value and is compared with the particular 'threshold' value to decide whether this sequence could be a 'site' or 'non-site'.

In addition, two algorithms for selection of the characteristics most informative for recognition are included in the program. In this case, the most informative characteristics for recognition are chosen based on the data on mutual correlations of the properties. Use of the weights calculated by the algorithms for selecting the most informative characteristics allows, in certain cases, for a decrease in recognition errors. Detailed information on the method is available in the Method Overview section of the web site. The calculation module is written as a program in C++; the web interface, in Perl.

INPUT AND OUTPUT

The resource SITECON is available at http://wwwmgs.bionet.nsc.ru/mgs/programs/sitecon/.

The standard way to use the web resource SITECON is as follows.

(i) Either select from menu 1 (Figure 1) a particular transcription factor binding site or use a user's sample (fields 2 or 3).
(ii) To analyze the conservative properties use button 5 to get the map and table of conservative properties.
(iii) For recognition, calculation of recognition errors (button 5) is necessary for determining the recognition threshold that is most suitable for the user. In the case of a user's own sample, window size and the algorithm for applying weights may be selected.
(iv) Input the chosen recognition threshold (fields 9 and 11) and the sequence to be tested (fields 6 or 7) and obtain the recognition result by clicking button 12.

In addition to the option 'user input', main menu 1 contains >100 variants of samples of binding sites for various transcription factors which can be used in the analysis of conservative conformational and physicochemical properties and recognition of sites. For this resource, high-quality representative samples have been constructed using the information compiled in the TRRD database (15). A site has been added to the corresponding sample only if either its functionality has been confirmed experimentally or if binding of the transcription factor to the site has been demonstrated by one of the following methods: EMSA (electrophoretic mobility shift assay) with nuclear extract and specific antibodies, EMSA with purified or recombinant protein, DNase I footprinting with purified or recombinant protein, or *trans*-activation of a reporter gene by overexpression of a transcription factor together with mutation analysis of site. In addition, we have used in this resource samples constructed of artificially selected sequences binding with high affinity to transcription factors from the database TRRD-ArtSite. These samples have been used to construct the knowledge base of conservative conformational and physicochemical properties of the corresponding transcription factors. The knowledge base can be accessed by selecting the corresponding sample from the list in the main menu and the option 'Map of conservative properties'. An entry in the knowledge base for a proposed sample is a map of conservative conformational and physicochemical properties (Figure 2); it is also available while performing

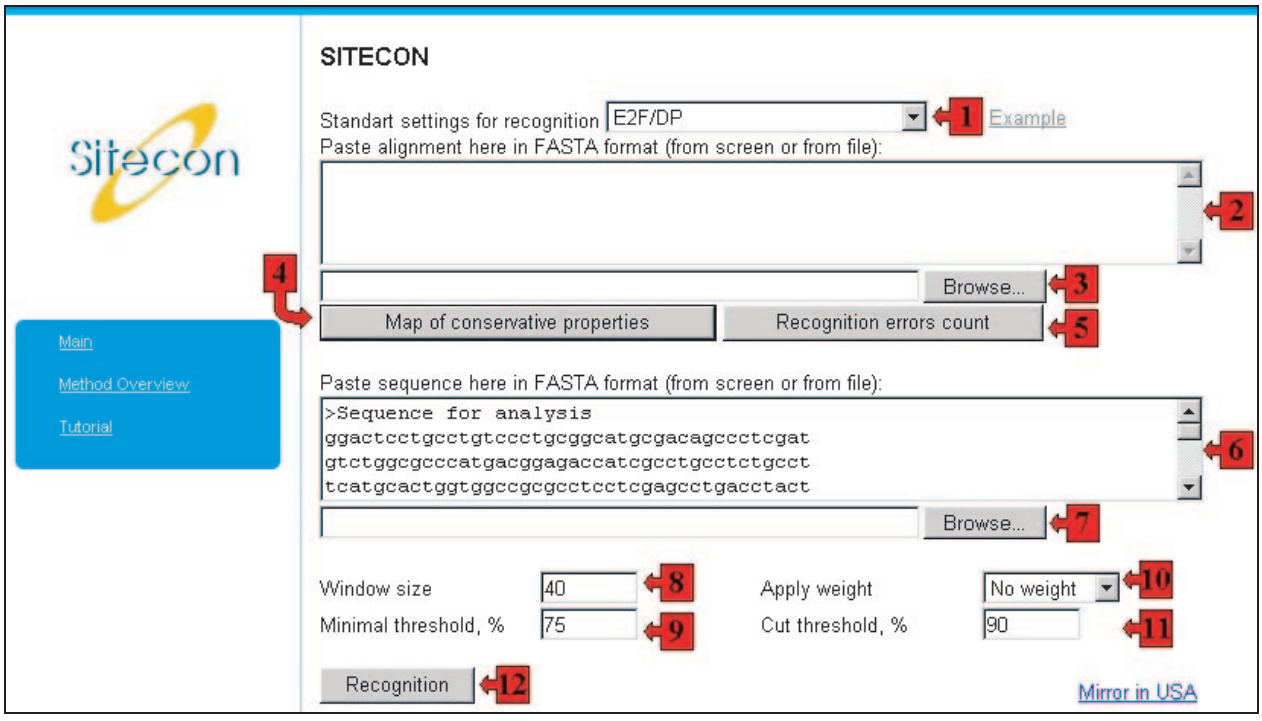

Figure 1. Interface of the web version of the program SITECON. The sample analyzed may be selected from the main menu of the program (**1**) or proposed by the user. In the latter case, the option 'User input' is selected and either a sample in FASTA format without gaps can be input into the window (**2**) or an alignment file can be selected on the user's PC using the option 'Browse' (**3**). A map of conservative conformational and physicochemical properties for the input sample is available on selecting the option 'Map of conservative properties' (**4**); and for calculation of recognition errors, the option 'Recognition errors count (**5**)'. The analyzed sequence or a set of sequences in FASTA format may either be input into window (**6**) or accessed using the 'Browse' (**7**) command from the PC of user. The bottom part of interface contains the block for inputting the four parameters for recognition, namely, size of recognition window (**8**), minimal recognition threshold (**9**) and (**10**) cutoff threshold for which the results will be output. It is also possible to select between three possible algorithms for calculating the recognition weights (**11**). Option (**12**) starts the recognition program.

recognition and testing recognition quality. The same map is calculated for a sample specified by user, when the same option is selected. In addition to the 'Map of conservative properties', a 'Table of 'conservative properties' is also available. This table contains information on the mean values and SD of properties for each position of the sample. This allows for analyzing and comparing individual properties selected by the user.

Recognition errors are calculated using another program, which helps to make a decision on the value of the cutoff threshold for the recognition program. Recognition errors at each threshold reflect the quality of the algorithm searching for sites. Type I errors shows the fraction of potential sites missed during the recognition. Type II errors allow for assessing the fraction of the sites recognized accidentally and, therefore, reflect the reliability that the site found is actually the site for binding of a given transcription factor. Choosing a low threshold will lead to the recognition of an excessive number of transcription factor binding sites with a low reliability. Choosing too high a threshold will lead to none of the transcription factor binding sites being recognized.

In the case of a user-specified sample, it is necessary before analysis to select the size of window and the algorithm for calculating the weight. Window size must not be smaller than the consensus sequence for this particular site and longer than the alignment length. It is optimal to choose a value of this parameter a little lower than the alignment length. When too big a window size is chosen, the program automatically reduces the value to the alignment length. Selection of the algorithm used for choosing weights is optional; in most cases, choosing no weights will suffice. In some cases, choosing any algorithm will increase the recognition quality. For the samples from the main menu, the optimal values of these parameters are fixed. In addition, the cutoff threshold may be corrected upon completion of the calculation by the recognition program; the value of the cutoff threshold is bounded below by the minimal threshold parameter.

In the output table of errors, calculation of an exhaustive set of standard measures of quality prediction is provided in addition to type I (false negatives) errors and II (false positives) for each threshold level of the required conformational similarity of potential sites. This list comprises correlation coefficient CC, averaged conditional probability ACP, sensitivity Sn, specificity Sp, second specificity coefficient SP, first prediction quality coefficient K_1, second prediction quality coefficient K_2, Yule's association coefficient Q, Ives and Gibbons correlation coefficient m and simple matching coefficient SMC (16,17). It is also possible to display recognition errors in graphical form.

In the output of recognition results, all the necessary information about the parameters of the recognition program, sites detected and sequences tested is available (Figure 3). The web

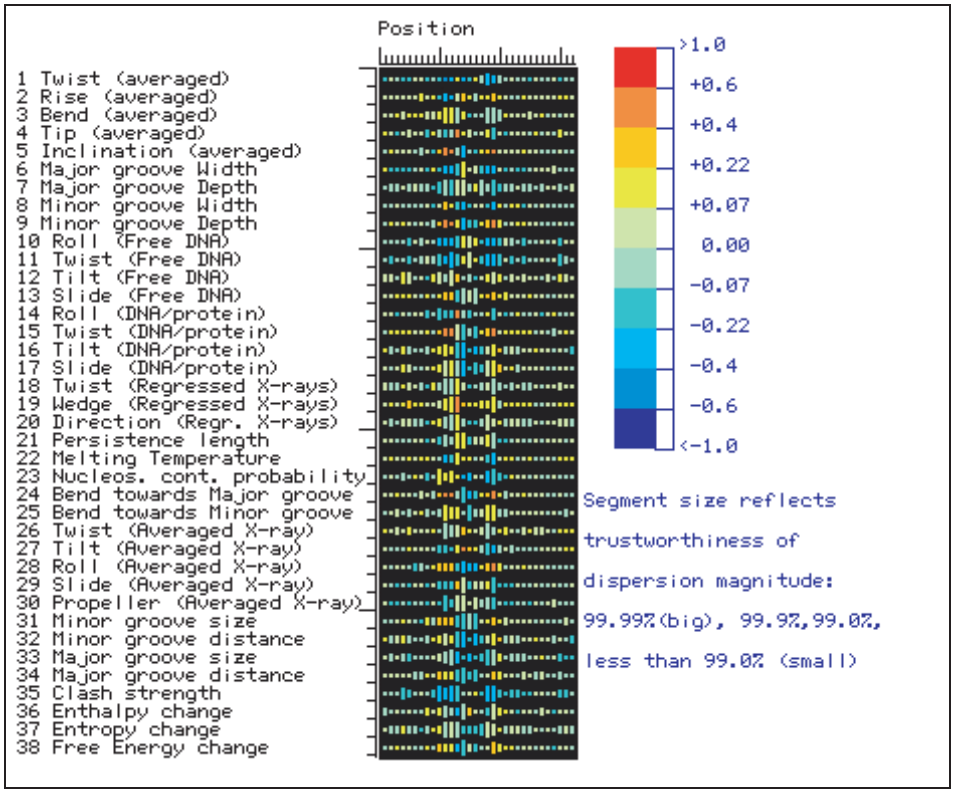

Figure 2. An example of the map of conservative conformational and physicochemical properties. The number of each property listed on the map corresponds to the number in the database 'Property'. The size of each colored block corresponds to the confidence level of the consistency of the hypothesis on conservativeness of a property. The color characterizes deviation of the mean value of a property from its mean value over the sample of random sequences and is calculated as the ratio of mean values of properties at positions of the alignment of the site sample analyzed to those of the random sample with the same nucleotide composition.

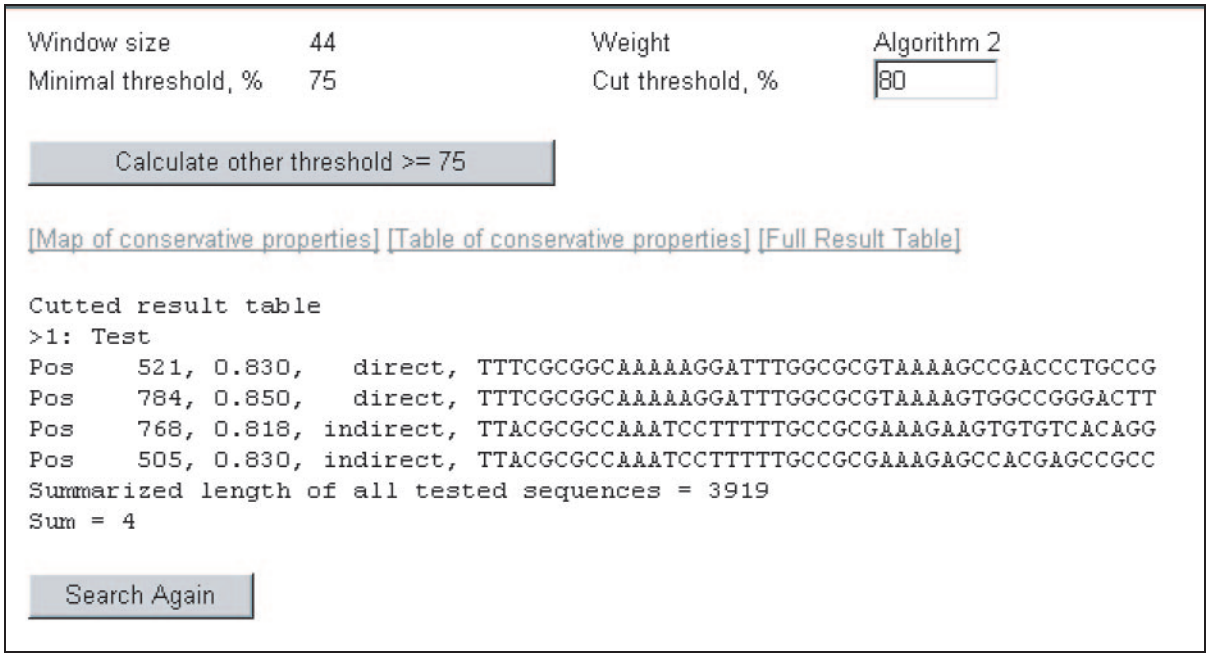

Figure 3. The output format of recognition results. Data on potential sites are preceded by the name of the sequence where they were found. Then, each line contains data on the potential site—position, level of required conformational similarity, orientation relative to the beginning of sequence and the sequence that was recognized as the site. For convenience, the total number of sites found in the sequences analyzed at a certain cutoff threshold and the total length of the analyzed sequences are given at the end of table. In the upper part of the results page are given the parameters of the calculation, references to the map and table of conservative properties, the full results table and the option to select a different threshold.

site includes examples of the analysis of transcription factor binding site samples, information about the algorithms used and a tutorial.

ACKNOWLEDGEMENTS

The authors are grateful to D. A. Afonnikov for fruitful discussion of the study and G. B. Chirikova for assistance in translation. The work was supported by the Russian Foundation for Basic Research (Grant nos 03-04-48469-a 02-07-90355, 03-07-90181-v and 02-07-90359); the Ministry of Industry, Science and Technologies of the Russian Federation (Grant no. 43.073.1.1.1501); the Presidium of the Russian Academy of Sciences (Grant on physicochemical biology no. 10.4); and NATO (Grant no. LST.CLG.979816).

REFERENCES

1. Starr,D.B., Hoopes,B.C. and Hawley,D.K. (1995) DNA bending is an important component of site-specific recognition by the TATA binding protein. *J. Mol. Biol*., **250**, 434–446.
2. Dickerson,T.D. and Drew,H.R. (1981) Structure of B-DNA dodecamer. II. Influence of base sequence on helix structure. *J. Mol. Biol*., **149**, 761–786.
3. Frank,D.E., Saecker,R.M., Bond,J.P., Capp,M.W., Tsodikov,O.V., Melcher,S.E., Levandoski,M.M. and Record,M.T.,Jr (1997) Thermodynamics of the interactions of Lac repressor with variants of the symmetric Lac operator: effects of converting a consensus site to a non-specific site. *J. Mol. Biol*., **267**, 1186–1206.
4. Suzuki,M., Amano,N., Kakinuma,J. and Tateno,M. (1997) Use of 3D structure data for understanding sequence-dependent conformational aspects of DNA. *J. Mol. Biol*., **274**, 421–435.
5. Meierhans,D., Sieber,M. and Allemann,R.K. (1997) High affinity binding of MEF-2C correlates with DNA bending. *Nucleic Acids Res.,* **25**, 4537–4544.
6. Oshchepkov,D.Y., Turnaev,I.I., Pozdnyakov,M.A., Milanesi,L., Vityaev,E.E. and Kolchanov,N.A. (2004) SITECON—a tool for analysis of DNA physicochemical and conformational properties: E2F/DP transcription factor binding site analysis and recognition. In Kolchanov,N. and Hofestaedt,R. (eds), *Bioinformatics of Genome Regulation and Structure*. Kluwer Academic Publishers, Boston/Dordrecht/London, pp. 93–102.
7. Liu,R., Blackwell,T.W. and States,D.J. (2001) Conformational model for binding site recognition by *E.coli MetJ* transcription factor. *Bioinformatics*, **17**, 622–633.
8. Ponomarenko,M.P., Ponomarenko,J.V., Frolov,A.S., Podkolodny,N.L., Savinkova,L.K., Kolchanov,N.A. and Overton,G.C. (1999) Identification of sequence-dependent DNA features correlating to activity of DNA sites interacting with proteins. *Bioinformatics*, **15**, 687–703.
9. Ananko,E.A., Oshchepkov,D.Y., Levitskii,V.G. and Pozdnyakov,M.A. (2002) Analysis of the regulatory regions of genes involved in the immune system operation. *Proceedings of the Third International Conference on Bioinformatics of Genome Regulation and Structure*, Russia, Vol. I, pp. 68–71.
10. Proscura,A.L., Levitsky,V.G., Oshchepkov,D.Y., Pozdnyakov,M.A. and Ignatieva,E.V. (2002) Expression of lipid metabolism genes: description in TRRD database and computer-assisted analysis. *Proceedings of the Third International Conference on Bioinformatics of Genome Regulation and Structure*, Russia, Vol. III, pp. 255–257.
11. Furman,D.P., Katokhin,A.V., Oshchepkov,D.Y. and Stepanenko,I.L. (2002) Do Drosophila retrotransposon LTRs contain functional sites capable of providing heat-shock-inducible transposition? *Proceedings of the Third International Conference on Bioinformatics of Genome Regulation and Structure*, Russia, Vol. I, pp. 89–92.
12. Turnaev,I.I., Oshchepkov,D.Y. and Podkolodnaya,O.A. (2004) Extension of cell cycle gene network description based on prediction of potential binding sites for E2F transcription factor. In Kolchanov,N. and Hofestaedt,R. (eds), *Bioinformatics of Genome Regulation and Structure*. Kluwer Academic Publishers, Boston/Dordrecht/London, pp. 273–282.
13. Ponomarenko,J.V., Ponomarenko,M.P., Frolov,A.S., Vorobyev,D.G., Overton,G.C. and Kolchanov,N.A. (1999) Conformational and physicochemical DNA features specific for transcription factor binding sites. *Bioinformatics*, **15**, 654–668.
14. Anderson,T.W. (1958) *An Introduction to Multivariate Statistical Analysis*. John Wiley & Sons Inc., New York, NY.
15. Kolchanov,N.A., Ignatieva,E.V., Ananko,E.A., Podkolodnaya,O.A., Stepanenko,I.L., Merkulova,T.I., Pozdnyakov,M.A., Podkolodny,N.L., Naumochkin,A.N. and Romashchenko,A.G. (2002) Transcription Regulatory Regions Database (TRRD): its status in 2002. *Nucleic Acids Res*., **30**, 312–317.
16. Baldi,P., Brunak,S., Chauvin,Y., Andersen,C.A. and Nielsen,H. (2000) Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, **16**, 412–424.
17. Bajic,V.B. (2000) Comparing the success of different prediction software in sequence analysis: a review. *Brief. Bioinform*., **1**, 214–228.